A SYSTEM AND METHOD TO ACCELERATE CLIENT/SERVER INTERACTIONS USING PREDICTIVE REQUESTS

Related Application

This application claims priority of U.S. provisional patent application Serial No. 60/183,818, filed February 22, 2000; U.S. provisional patent application Serial No. 60/194,050, filed April 3, 2000; and U.S. provisional patent application Serial No. 60/196,163, filed April 11, 2000; each of which are hereby incorporated by reference in their entirety.

Field of the Invention

The present invention relates to the field of data communications. More specifically, the present invention relates to the enhancement of perceived data throughput in a server/client communications system.

Background of the Invention

Data communications systems based on a dispersed client/server architecture have become prevalent over the last decade. The most notable and extensive of these dispersed client/server communications systems is the Internet. The World Wide Web (the "Web"), a series of applications introduced in the early-to-mid 1990's which allow layman Internet users to access information residing on servers at all points of the globe, has caused the volume of Internet data traffic to increase to unprecedented levels. The explosive growth in Internet data traffic, which seems to have outpaced the rate of expansion of the Internet's infrastructure, has produced bottlenecks along many of the Internet's data pathways. These bottlenecks intern cause delays in data flow between users and the servers with which they attempt to communicate.

Summary of the Present Invention

As part of the present invention, a Client Agent receives or intercepts a request for data from a client application running on a user's computer. The Client Agent may analyze the request and may forward the request to a server having the requested data or to a Predictive Server associated with the server. The Predictive Server may analyze the request and forwards the request to the server. A response by the server to the request may be intercepted either by the Predictive Server, the Client Agent, or both. Either the Client Agent or the Predictive Server may generate one or a series of predictive requests for data to the server based on the content of the server's response. A response by the server to a predictive request may be stored either at the Client Agent or at the Predictive Server, and may be transmitted to the client application when the client application transmits a request for the data contained in the response. In another mode of the invention the Client Agent causes the client to accelerate his requests. This mode can be used separately or in conjunction with other modes.

In the provisional patent applications (60/183,818; 60/194,050 and 60/196,163) the Client Agent is referred to a "Predictive Agent".

Brief Description of the Drawings

The present invention will be understood and appreciated more fully from the following detailed description taken in conjunction with the appended drawings in which:

Fig. 1 is a diagram showing a Client Agent and a Predictive Server according to the present invention;

Fig. 2 is a data flow diagram illustrating a first mode of the present invention;

Fig. 3 shows a hyper-text-markup-language ("HTML") web page having URL's of the objects contained within the page;

Fig. 4 is a diagram illustrating compression of data on the channel between a Client Agent and a Predictive Server;

Fig. 5 is a data flow diagram illustrating a second mode of the present invention; and

Fig. 6 is a data flow diagram illustrating a third mode of the present invention.

Detail Description of the Invention

The present invention is a system and method of enhancing perceived data throughput between a client application and a server containing data to be retrieved by the client. Predictive strategies implemented by intermediary elements such as a Client Agent and/or a Predictive Server, cause the server to transmit data likely to be requested by the client application, in advance of the client actually requesting the data or causes the client to accelerate his requests. The data may be stored either at the Client Agent or the Predictive Server, and is transmitted to the client application when a request for the data is actually transmitted by the client. Either the Client Agent or the Predictive Server may act as a predictive unit, that is, as a source of a predictive request. Either the Client Agent or the Predictive Server may act as a buffer unit, that is, as temporary storage for a server response.

Requests for data and responses thereto may be transmitted in data packets or other common data structures used in connection with a dispersed data network. TCP/IP and/or any other communication protocols may be used as part of the present invention. The data transmitted using the system and method of the present invention may be text, graphics, audio or video. Requests for data may be transmitted using

Hyper Text Transfer Protocol ("HTTP") or any other data retrieval and exchange standard, currently known or yet to be devised.

Turning now to Fig. 1, there is illustrated a system according to the present invention, wherein a Client Agent 100 and a Predictive Server 200 are located along a data pathway between a client and a server. Both the Client Agent 100 and the Predictive Server 200 have a storage unit, 120 and 220 respectively, and both have an analyzer unit 110 and 210 respectively.

The Client Agent 100 may be a software application residing on a computer in proximity with the computer where the client application is running, or even on the same computer as the client application. The Client Agent 100 may be an application also functioning as an Internet proxy for the client application.

The Predicative Server 200 may be an application running on a computer in proximity with the server, or even on the same computer as the server application.

A request for data from the server, generated by the client, may follow a path from the client, through the Client Agent 100 (path 1), through the Predictive Server 200 (path 2) and to the server (path 3). Alternatively, the request may follow a path skipping either the Client Agent 100 (path 4) or the Predictive Server 200 (path 5), but not both. Although the Client Agent 100 or the Predictive Server 200 may not be a recipient of a request generated by the client or a response generated by the server, either may intercept and analyze a copy of the communication. The Client Agent or Server may analyze data packets passing by using a method known as sniffing the line, or any other equivalent method, and intercept those packets having a destination address value related to either the client or the server.

A response from the server may follow a reverse path from that of the request evoking the response, or any other path illustrated in Fig. 1. Although a Client Agent

100 or the Predictive Server 200 may not be a recipient of a response generated by the server, either may intercept and analyze a copy of the request.

The transmission of a client's request or a server's response is regulated via the manipulation of source address and destination address values on the data packets comprising the data object, or by any other means presently known or not yet devised. For example, a data packet transmitted by a client and addressed to a server may be intercepted by either the Client Agent 100 or the Predictive Server 200, and re-addressed to have the destination address of a destination other than the server and a source address of the Agent, which intercepted the packet. Thus, altering the path of the packet. The routing and/or re-addressing of data packets between different points on a dispersed network are well known. The present invention functions with all known and not yet devised methods of routing and/or readdressing data packets.

The system and method of the present invention may take one of several embodiments and may be implemented according to one of several methodologies, examples of which are described below:

Mode One – Client Agent & Predictive Server:

Turning now to Fig. 2, there is illustrated a data flow diagram for a first mode of the present invention where both a Client Agent 100 and a Predictive Server 200 are utilized. As part of the first mode of the present invention, a client's request for data 310 is received by the Client Agent 100, where a record of the request is made. The Client Agent 100 then forwards the request 320 to the Predictive Server 200, where a record of the request is also made. The Predictive Server 200 then forwards the request 330 to the server from which a response is expected. The above-mentioned steps may be collectively referred to as Stage 1, as shown in the Fig. 1.

Stage 1 ends and Stage 2 begins upon the server's receipt of the client's request. After analyzing the request and assuming the server possesses in its storage device the requested data or web page, it generates and transmits a response 340 corresponding to the request. The server's response travels back through the Predictive Server 200, where it is analyzed to determine one or a series of possible subsequent requests the client may transmit. Typically, one of the first responses by a server to a client's request for a web page contains a list of objects present within the page, and instructions as to how to retrieve these objects from the server (e.g. the objects' URLs). The Predictive Server 200 may derive the series of possible future client requests and generate a "prediction list" based on the list of page objects in the response. Fig. 3 shows a hyper-text-markup-language ("HTML") web page having URL's of the objects contained within the page. The Predictive Server 200 forwards the response 350 to the Client Agent 100 and issues one or a series of predictive requests 331, 332...etc.., shown in Fig. 2 as dashed arrows, to the server. The predictive requests are based on the results of the analysis of the server's previous response. In the case where the original request is for a web page, the series of predictive requests will typically be URL's of the objects contained within the page.

In response to the predictive requests, the server may issue predictive responses 341, 342.. addressed to the Predictive Server 200. The Predicative Server 200 may automatically forward the predictive responses to the Client Agent 100, or the Predictive Server 200 may store the predictive responses in storage 220 and wait to receive a request from the Client Agent 100 for a specific response or set of responses before forwarding it. In the example of Figure 2, when a Predictive Server 200 receives any response from the server, it automatically forwards the response to the

Client Agent 100, in particular the first response 350 is immediately forwarded to the Client Agent.

Upon receiving the first response 350 from the Predictive Server 200, the Client Agent 100 forwards the response 360 to the client and generates its own "predictive list" by performing a similar analysis on the first response 350 to that which was performed by the Predictive Server 200 when the Predictive Server 200 generated its "predictive list".

The client, upon receiving a first response 360 from the server, compares the list of objects within the response against objects already stored locally, and issues a request 311 which it forwards to the Client Agent 100. The Predicative Agent 100 compares the request 311 against its own "predicative list" and against a list of already received predictive responses. If the predictive list does not have an entry for a predictive request corresponding to the client's request 311, the Client Agent 100 forwards the request along, as shown by arrow 325 in Fig. 2. If a predictive response corresponding to the request 311 is on the predictive list and has already arrived at the Client Agent 100, the response is transmitted to the client, as shown by arrow 361. If a corresponding predictive request is on the list, but no corresponding predictive response has yet arrived at the Client Agent 100, the Client Agent 100 waits for the corresponding response to arrive and forwards it to the client upon receipt.

Fig. 4 shows a mode of the present invention where the Predictive Server 200 does not automatically forward to Client Agent 100 predictive responses received from the server 341, 342, as shown in Fig. 2, but instead stores the predictive responses in storage 220 until a request 421 for the responses is received from the Client Agent 100. The client, upon receiving a first response 360 from the server,

compares the list of objects within the response against objects already stored locally, and issues requests 411, 412, ..414 for only those objects not present locally. The client's set of issued requests 411, 412, ..414 may have fewer objects than contained in the response 360 when one or several of the objects listed in the response are already present locally at or near the client. The Predicative Agent 100 then forwards the set, or an equivalent representation of the set 421, to the Predictive Server 200. Thereby, the Predictive Server 200 only transmits those predictive responses 451, 452..454 not already present locally at the client. Upon receiving the responses, the Client Agent 100 transfers them to the client 461, 462...464.

As part of the present invention, the Client Agent 100 and the Predictive Server 200 may perform communication optimization techniques such as compression, protocol conversion, and partial information transfer on the connections (i.e. requests and responses) between the two. Many techniques and strategies are known and applicable to the present invention. For example, the Client Agent 100 may combine several client requests into one request packet or a series of packets, as shown in Fig. 4. Additionally, the Client Agent 100 may convert a request containing a complete Uniform Resource Locator ("URL") for an object into a smaller request with only a partial or relative URL. That is, the smaller URL contains only part of the instructions for retrieving an object from the server. However, because the Predictive Server 200 is able to identify the source of the smaller request, namely the client, and the Predictive Server 200 has a record of the last response received by the client, the Predictive Server 200 is able to convert a request with a partial URL back into one with a complete URL which is in a form acceptable to the server.

Responses to client requests, transmitted from the Predictive Server 200 to the Client Agent 100 may also be compressed by combining several responses into one or

a series of packets. Various compression routines may be applied data flowing from the Predictive Server 200 to the Client Agent. Upon receipt of compressed data, the Client Agent may decompress the data and forward it to the client in its original form. Compression and decompression of data is well known, and any methods known today or yet to be devised may be applied as part of the present invention.

In an alternate embodiment of the present invention, the predictive responses are addressed directly to the Client Agent 100, thereby bypassing the Predictive Server 200. Each of the predictive requests sent by the Predictive Server 200 may contain a source address of the Client Agent 100. When a predictive request contains the source address of the Client Agent 100, the server's response to the predictive request, the predictive response, is addressed directly to the Client Agent 100.

The Client Agent's 100 "predictive list" of requests may be derived from the response, in the same manner as a list was derived by the Predictive Server 200, or the list may be a duplicate of the list produced by the Predictive Server 200. There are many well-known methods by which the Predictive Server 200 may transmit its prediction list to the Client Agent 100. The present invention may utilize any method or protocol of transmission, currently known or yet to be devised.

Mode Two – Client Agent

Turning now to Fig. 5, there is shown a data flow diagram for a mode of the present invention only utilizing a Client Agent 100. As part of the illustrated mode, a client's request 510 is received by a Client Agent 100 and sent 520 directly to a server. The server's response 530 to the request is intercepted by the Client Agent 100 and stripped of all information other than page formatting and the list of objects needed to be retrieved in order to complete the page, with a command to re-load all the objects after they are all retrieved, by using for example a Java Script. When the client

receives this modified and stripped down version of the response 540, it checks against a list of locally stored objects to determine which objects need to be requested. The Client issues requests 511a, 512a ..514a for those objects not present locally. The Client Agent 100 forwards the request 521-524 to the server and responds to each of these requests with a pseudo or fake response 551-554 containing little or no data.

While the client is receiving pseudo responses to its requests, the server is sending real responses to the request 531-534 to the Client Agent 100. Once the client runs the reload script, the Client Agent 100 receives a new set of requests 511b-514b from the client. The new set is just a copy of the previous set 511a-514a, and the Client Agent 100 checks each incoming request for matching or corresponding response, which may have already arrived from the server. In the event a response has arrived, the Client Agent 100 forwards the response 541-544 to the client. If a corresponding response has not arrived, the Predictive Server waits and forwards the response to the client as soon as it is received.

This mode of the invention may be practiced utilizing only a Predictive Server 200 instead of a Client Agent 100.

Mode Three – Predictive Server

Turning now to Fig. 6, there is a data flow diagram illustrating data flow of mode of the present invention where only a Predictive Server 200 is used. As part of this mode, the Predictive Server generates or issued of series of predictive requests 651-653 based on a server's first response 630. The first response 630 is forward to the client 640 which checks to see which of the objects listed in the response are present locally, and then the client issues requests 611-614 for those objects listed in the response but not present locally. The requests 611-614 are received by the

Predictive Server 200 and corresponding responses are sent to the client as soon at they are received by the Predictive Server 200. Given that the Predictive Server 200 issued predictive requests 651-654 at about the same time as the original response 610 was forwarded to the client, many of the responses for the requests 611-614 issued by the client should already be received 631-634 and be stored at the Predictive Server 200 storage 220 prior to the receipt of the client's requests 611-614.

It is appreciated that one or more steps of any of the methods described herein may be implemented in a different order than that shown while not departing from the spirit and scope of the invention.

While the methods and apparatus disclosed herein may or may not have been described with reference to specific hardware or software, the methods and apparatus have been described in a manner sufficient to enable persons of ordinary skill in the art to readily adapt commercially available hardware and software as may be needed to reduce any of the embodiments of the present invention to practice without undue experimentation and using conventional techniques.

Those skilled in the art will appreciate that the present invention can be used in any client/server architecture, which architecture uses a systematic method in which the client retrieves data from the server (e.g. MS Exchange, Lotus Notes, SAP,.. etc.).

While the present invention has been described with reference to a few specific embodiments, the description is intended to be illustrative of the invention as a whole and is not to be construed as limiting the invention to the embodiments shown. It is appreciated that various modifications may occur to those skilled in the art that, while not specifically shown herein, are nevertheless within the true spirit and scope of the invention.